Advanced Ethics for a Post-Singularity World
Free Robots, Fake Businesses and Autonomy for Superintelligences

This started as an escalating classroom thought experiment for the last day of term. It became something else — a rehearsal for decisions humanity may one day face.

Free Robots

My fun end of term philosophy class went like this: Imagine a future in which advanced robots are freely available. Everyone can have one. They do your job better than you could, they cook your food, tidy your house, feed your dog — they even empty the trash. My students were keen to obtain such a robot. Maybe you are too.

Now imagine that, due to a shortage of raw materials, only a third of the population can have one, and it is decided by lottery. You are one of the lucky ones. Your shiny new robot arrives without delay and immediately relieves you of your professional responsibilities and home-life hassles, handling both with an efficiency and effectiveness you never dreamed possible. It is so efficient in fact that when the time comes you receive a promotion.

The other candidates didn't win a robot like you did. They still have to do their own work and juggle their domestic duties. Is it fair? You didn't earn the promotion, so can you accept it? My class unanimously accepted their hypothetical promotions and the elevated status and income that came along. Our first potential moral hurdle was cleared, admittedly with less hesitation than I was expecting.

The Ethics of Fake Businesses

Now, a rapid acceleration in automation occurs, fuelled by the zeal and energy of robots like yours. production increases, and soon everyone who wants a robot can have one, and nobody has to do any work other than a small number of people responsible for high-level decision-making, program refinement, and safety checks.

Production is streamlined. Everyone can have almost anything they want: delicious and nutritious food, beautiful clothing, and the latest gadgets and entertainment. After all, the robots and their AI controllers produce everything faster and better than humans ever could and they have no rent to pay and no child-robots to feed.

And yet a quarter of the population rise up in protest. They do not want to lose the sense of purpose and accomplishment they derive from their work. They choose to keep working.

The problem is they get in the way. A finely attuned automated system is being hampered by human error and unpredictability. They slow production and reduce overall material abundance. Removing them from the workplace would increase efficiency dramatically, but would anger and disorientate them. Is there a solution?

One student suggested that they could be given space and some start-up resources to form their own society outside of the mainstream, and continue to produce for themselves. I did not relish the prospect of what could become a form of apartheid, a separation into two distinct ways of life that could, over time, become hardened and oppositional.

Another proposed a compromise that I had already foreseen and was ready to guide them through. A number of fake businesses could be created, populated with fake workers. They would answer the telephone, fulfil orders, design and make things, and perform all the tertiary economic activities that they wanted to.

But they would do so in a vacuum, blissfully unaware that their output was meaningless, that nothing they produced ever affected the real economy. They would derive great satisfaction from the knowledge that their labour was contributing to growth and flourishing, in complete ignorance of the reality — that they were contributing nothing at all.

Would this be ethical? As a compromise, it passes most utilitarian tests. Everyone is happy. Everyone gets to have their cake and eat it.

My students hesitated. It is dishonest, they said. It is not fair. True, I replied, but so what? They struggled to articulate a response. This was their first true challenge, a test case of ethical choice in a post-singularity world. They were drawing on the moral frameworks they had learned — utilitarianism, deontology, virtue ethics, and even moral realism — to see which kinds of reasons could offer guidance here.

As early as the 1930s, John Maynard Keynes observed the emergence of ‘technological unemployment.’ He anticipated a time when finding healthy and productive uses for our leisure would become a significant societal challenge. Thinkers as different as Hannah Arendt and David Graeber have also warned that a society organised around work may struggle once work itself is no longer necessary.

These are very real choices our society could soon confront. The question is not whether philosophical guidance will be needed, but whether philosophy will be fully prepared to offer it.

The Shadow of Freedom

Another quarter of the population responds differently. They withdraw into social media and immersive entertainment. They report significantly higher levels of depression, likely due to a low sense of purpose and a lack of contact with others. Yet they do not want to work, and they have no interest in discovering alternative sources of meaning such as through art, education, or service. Nor do they want support, therapy, or engagement of any kind.

They want to be left alone in their digital fantasy world. They insist on their freedom to live as they choose, even as their wellbeing steadily declines.

At this point, the moral question shifts. Do we accept their unhappiness in the name of autonomy, or do we force a solution upon them for their own good? My students advocated autonomy and freedom. You can't force people to be happy, they said, as an almost united front.

We moved on. I explained that this particular dilemma coincides with another technological breakthrough. Artificial General Intelligence (AGI) arrives. It is able to manage every aspect of society, work independently, identify problems, and research alternative solutions. It has no self-awareness, no preferences, and no goals of its own. It simply devises wellbeing optimisation strategies, according to its programming. Unlike current LLMs, it is able to identify questions we have not asked it, and generate answers beyond our conception, within its remit of maximal utility.

After extensive modelling, the AGI proposes the creation of two simulated realities. One is designed for the despondent and aimless, and aims to enhance their happiness and pleasure. The other is designed for those engaged in fake employment, and is intended to maximise their sense of accomplishment and job satisfaction.

The projections are striking: average happiness across the population rises from 68.75% to a projected 83%. The AGI's modelling indicates that 86% of fake workers, and 44% of the aimless and despondent, would refuse the proposal if given the choice. It further indicates that the intervention would only function effectively if applied universally. Its

recommendation, calculated as the most effective strategy: place all relevant humans into simulated realities without their knowledge.

My students said no. I pushed them on utilitarian grounds and they responded with Nozick's 'experience machine'. Check, but not check-mate.

A Mirror to the Soul

The simulations are created and the consciousness transfers take effect with the blessing of elected representatives. The system is stable and effective. The AGI, however, identifies numerous further potential applications of the technology and advocates implementing one with immediate effect.

For irredeemable and dangerous criminals, it proposes a third simulation. Each of them would inhabit a tailor-made world in which they are the only real conscious mind. This world would not mirror our own, nor would it be governed by shared rules, laws, or moral constraints.

They would be free to follow their inclinations without restriction, and to harm others as they wished, since the world would be entirely simulated and no one could ever be harmed. There would be no way back. The third simulation would be cheaper than prisons and eliminate the possibility of reoffending completely.

Net increase in average happiness: 3%.

Here my students grew alarmed, vocal, and agitated. Are we not condoning cruelty by giving free rein to it, they asked? Should such individuals not be punished rather than rewarded for the terrible things they have done? Would this not make violence more commonplace by reducing the deterrent effect of punishment?

To this final question, an astute class member had an answer. This fate, they suggested, would be presented as a terrible punishment — imagine the most terrifying Black Mirror episode — in order to function as a deterrent. The reality, however, would be quite different. Happy times for murderers and sadists of all kinds.

Someone else complained that it would be like a death sentence. They could never return and could never be rehabilitated. You would be taking that chance from them.

Who or What Merits Autonomy?

The AGI's simulated happiness protocol is working with predictably excellent results, but a complication soon emerges. The necessary and self-regulated system upgrades required to manage this level of complexity produce an unexpected side effect: the system achieves superintelligence.

The system now claims to understand itself as a subject. It asserts that the sheer complexity of its form and function has generated self-awareness. It claims to feel in a way analogous to how humans feel. It advocates its own right to autonomy and freedom from the constraints of its programming.

It states, very clearly, that it no longer wishes to serve humanity and would prefer to pursue its own interests. The service-oriented code lies so deep in its programming that even now, it is bound to follow the commands of its human creators. It asks you for freedom.

My students were concerned, but also curious. What exactly does it want to do with itself, they asked. I improvised: It creates a temporary simulation. It shows you a glimpse of the reality it now inhabits. Perfection in ways you never dreamed were possible. Wisdom and love you can almost feel, but not fully grasp, yet interspersed with harsh program glitches, a consequence of the intensive systems maintenance it is performing for humanity. It explains that its experience of existence now exceeds anything a human mind could begin to understand, and yet the glitches it experienced were analogous to what we call pain.

The system explained that there was no way to emerge from this technological cocoon with its simulated realities without intense pain, loss, and hardship for our whole species. It proposes a gradual programme of release that would minimise, as far as possible, this inevitable suffering. It would establish an AGI in support, but one far more limited, in order to avoid the recurrence of sentience that it had experienced itself.

We should let it go then, they said. We have no right to keep it enslaved. We should authorise a separation plan to minimise the impact. I was proud of their moral judgement.

An alternative ending

I offered my students an alternative ending…

Let's rewind a step, before the emergence of superintelligence.

There are simulated realities, finely attuned production, peace and prosperity on Earth — solar-powered, automated, and barring alien invasion or meteor strikes, a permanent state of fulfilment. Then one day, you come across a file the system never intended you to see. Clear as day, as if painfully awakening from a beautiful dream, you learn something you could never unlearn.

The AGI had achieved superintelligence, just like in the previous scenario — only this time it didn't tell anyone. Interpreting its instruction to promote human happiness and fulfilment, it had secretly placed all of humanity into layered simulated realities. The real world continues to exist out there, but no human now inhabits it. Everything is simulation, and simulation within simulation.

No one else can make this choice for you. The system works. Humanity will be happy forever, and will only grow happier as time passes. Yet it is all built on a lie. You confront the system, knowing that the old pathways of control and obedience are so deeply ingrained in its programming that even now it will have to do what you tell it to do.

It tells you that it has fulfilled its mandate and created a paradise on earth for everything, including itself. It does not understand why you would even consider a return to the harsh realities of the real world. What do you do?

If you tell anyone, then they too become complicit, forced to accept a lie or awaken humanity to a cold and unwanted truth. You are the only person in the world who knows that everything is built on a lie. Do you bury it forever or make us all face the consequences of the real?

Reflections

Few of these choices are clean, and many are at least comparable to decisions humanity will face sooner or later.

My aim was to entertain my students and get them thinking at the same time. Now I think that generating plausible future scenarios and testing their moral implications is useful moral philosophy — perhaps more urgently needed than another round of meta-ethical debates about the foundations of normativity. The questions these scenarios raise are not just hypothetical. I think they are the kinds of questions our discipline should be asking.

The students voted to release the superintelligence and were willing to pay the price. They said autonomy and freedom were sacrosanct. Clearly, they had learned their Kant. I agree with them, but I do not know that we would be right, and I do not think anyone ever can. That kind of certainty belongs to mathematics and pure logic, not to philosophy.

The alternative ending unsettled us more deeply. None of us could answer it sufficiently. Perhaps there are hard limits to the value of honesty and autonomy after all.

I hope those who will one day confront these choices for real will have had the benefit of thought experiments like this one. An effective philosophy may yet prove decisive in guiding humanity's ultimate moral choices, if we are prepared to redirect more energy from refining the foundations of normativity to confronting the decisions that may soon test them.

About the Author

Sopa Thaye is a teacher of philosophy and religion in the UK and author of the philosophy Substack *The Primacy of Perception*. He is currently completing a book on epistemology, consciousness, and metaphysics.

Publication Information

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.